

# Weekly Report

January 7, 2018

## 1 Work

### 1.1 降维

为了处理在大数据上使用Kmeans耗时的问题（虽然每次迭代是线性的，但是随着数据量增加，迭代的次数也要增多，所以总体来说耗时还是非线性的）。本周我们采用了四叉树的方法对二维投影进行划分。降维的效果和之前采用kmeans的方法差不多，但是时间可以从30秒降低到0.5秒。

再此基础上还有一个想法是对于四叉树的一个Cell中的点，基于knn graph进一步划分。因为一个cell中的点可能是两个label的数据，不应该一起移动，划分之后会更有道理。写了一个周末的程序发现：基于我们目前10NN或者100NN的Graph，同一个Cell中的点会被分为非常多的小类（70000的数据点生成4万个类，这对于优化没有帮助），也就是说直接数据点之间基于KNN Graph的联系还是太弱。除非计算更大的KNN Graph或者再把小类做进一步聚合，才能解决类数目过多的问题。然而，这样的计算代价过高，有可能会接近Kmeans。所以，我们目前还是直接采用四叉树划分的方法。

同时，本周继续在晚上论文的内容，计划下周出一个论文的初稿。

### 1.2 工作进度

Table 1: 工作进度

TASK	PROGRESS	DATE
dimension reduction	1)节省内存；2)写论文	1.10
location2vec专利		
*2Vec survey		1.30

## 2 Paper Reading

### 2.1 Optimizing F-Measures by Cost-Sensitive Classification

本文讨论了机器学习中衡量函数 $F_\beta$ 和机遇Cost-Sensitive为目标函数之间的联系，证明了两者的最优值是一样的。

### 2.2 Bubble Treemaps for Uncertainty Visualization

文章提出了一个更加紧凑的Bubble Treemaps的可视化方法，并且将不确定性编码在边界线上。经典的locally linear embedding算法，主要思想是一个点可以由他周围点的线形表达构成。我们的目标是使得重构的误差要在低维空间中比较小。

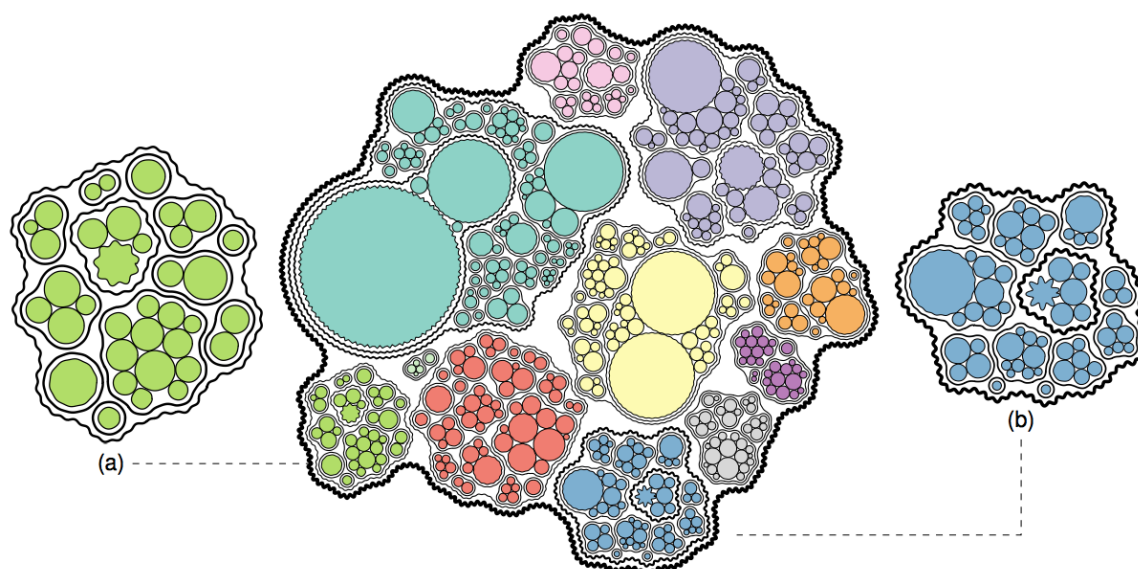


Figure 1: Bubble Treemaps

### 2.3 Visualizing Social Media Content with SentenTree

社交网络上的文字都是以句子的形式出现，所以普通的词云只能显示单词的频率不能展现句子的内容。本文提出SentenTree的方法，同时展示语义结构和单词频率。这样的思想应该可以扩展到文章的内容（对象是更长的文章），或者网络中的社团，对其主要信息作提取，并且展示相关上下文信息。



Figure 2: SentenTree

## 2.4 Patterns and Sequences: Interactive Exploration of Clickstreams to Understand Common Visitor Paths

一篇对于时序数据（网页点击数据）的多个层次聚合方法的文章。